Theory-Inspired Deep Multi-View Multi-Label Learning with Incomplete Views and Noisy Labels

Supplementary Material

In this supplementary file, we provide more details of our method and more experimental results from the following aspects: 1) Related Work; 2) Experimental Setup; 3) Comparative Experiment; 4) Parameter Determination, Convergence and Time Efficiency Analysis; 5) Visual Explanation; 6) Theoretical Derivation and Proof.

A. Related work

A.1. Incomplete Multi-view Multi-label Learning

Initial efforts to address incomplete data in MvMLC involved the exploration of many traditional methods. iMVWL [17] seamlessly integrated the learning of a shared subspace and weak-label classifier by leveraging label correlations and cross-view feature relationships. TM3L [21] utilized subspace learning and view weighting for feature extraction, combined with label completion and kernel extreme learning machines for efficient two-step optimization. NAIM3L [7] implicitly aligned non-aligned views in a shared label space while incorporating the global high-rank and local low-rank multi-label structures. However, traditional shallow models have faced challenges in effectively capturing intricate feature semantics and fine-grained correlations between labels. In contrast, recent advancements in deep learning based methods have demonstrated exceptional performance. DIMC [18] employed view-specific deep feature extraction and weighted fusion classification modules to mitigate the negative effects of missing data. DICNet [9] proposed deep instance-level contrastive learning to align views and utilized a missing-label indicator matrix to filter out invalid labels. VIST [14] utilized a view-category interactive sharing transformer, missing view generation, and embedding consistency enhancement to achieve efficient complementary fusion of views and labels. MTD [11] employed a two-channel decoupling framework to divide view features into shared and proprietary parts and integrated random fragment masking with label-guided graph regularization. AIMNet [10] introduced an attention-induced missing instance imputation technique, along with a multi-view late fusion strategy and label semantic feature learning. SIP [12] compressed cross-view representations to maximize shared information, while modeling label prototypes in the latent space.

A.2. Learning with Noisy Labels

Previous studies have thoroughly investigated the area of single noisy label learning. The proposed methods include the development of robust loss functions, noise transition matrix based approaches and label rectification with inter-class relationships. Robust loss methods focused on designing loss functions to mitigate the impact of noisy labels, such as Mean Absolute Error (MAE) [3] and Generalized Cross Entropy (GCE) [20]. [13] presented a loss normalization technique called Active Passive Loss (APL), which synergistically combined two robust loss functions that mutually boost each other to optimize training efficacy. [15] utilized a noise transition matrix to summarize the probabilities of one class being mislabeled as another, allowing for loss correction in the objective function, while VolMinNet [8] proposed an end-to-end framework without anchor points by minimizing the volume of the transition matrix and optimizing the cross-entropy loss. [4, 5] considered the complementary relationships between labels by leveraging a negative learning strategy. Class-conditional multi-label noise has recently become a focus of exploration since researchers actively pursue effective approaches. [19] developed two unbiased estimators for classifer learning under multi-label noise, while the reliance on a linear model limited its applicability to complex multi-target tasks. UNM [1] utilized a combination of cyclical learning rates and loss ranking to identify noise. [16] introduced a curriculum learning strategy to progressively identify true labels in the candidate set. These methods necessitate strong dependence on strategy adjustments, which leads to relatively insufficient adaptability and stability. ENTMLC [6] leveraged label correlation to estimate the noise transition matrix. The disjoint process of estimating the transition matrix and training the classifier results in cumulative errors, undermining its practical applicability.

B. Experiment

B.1. Experiment Setup

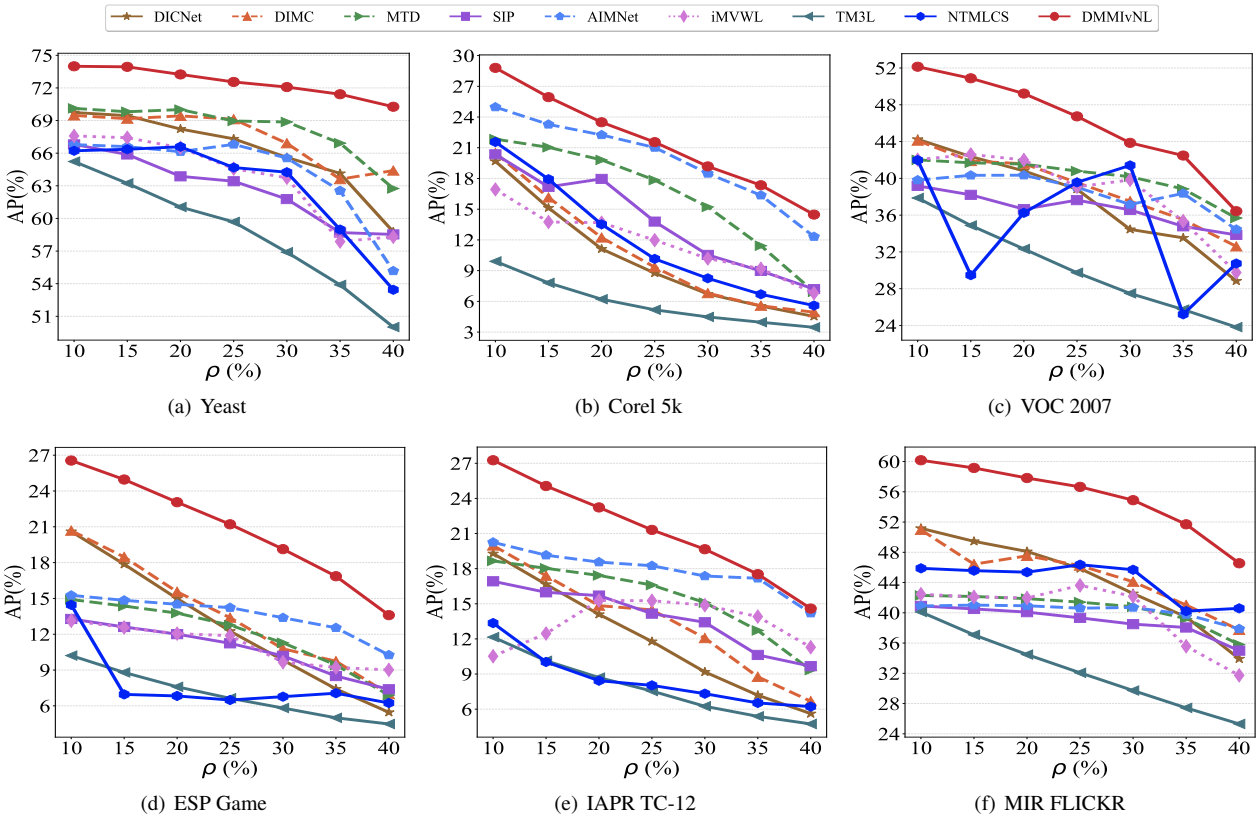
Datasets and Comparison Methods. In our experiments, six public multi-view multi-label datasets are selected as shown in Table 3. Their specific descriptions are as follows. **Corel 5k** is composed of 4999 image samples and 260 words, where each word can be regarded as an annotation or label. **IAPR TC-12** comprises 19627 high-quality natural images and each image contains 261 labels, including sports, actions, animals, cities, and so on. **ESP Game** is a multi-view multi-label dataset containing 20770 images with 268 corresponding tags. **VOC 2007** is a widely utilized dataset for visual object detection and

Table 1. Detailed information of datasets.

View	Yeast	VOC 2007	Corel 5k	Esp Game	IAPR TC-12	MIR FLICKR
1	Genetic Expression(79)	DenseHue(100)	DenseHue(100)	DenseHue(100)	DenseHue(100)	DenseHue(100)
2	Phylogenetic Profile(24)	DenseSift(1000)	DenseSift(1000)	DenseSift(1000)	DenseSift(1000)	DenseSift(1000)
3	-	GIST(512)	GIST(512)	GIST(512)	GIST(512)	GIST(512)
4	-	HSV(4096)	HSV(4096)	HSV(4096)	HSV(4096)	HSV(4096)
5	-	RGB(4096)	RGB(4096)	RGB(4096)	RGB(4096)	RGB(4096)
6	-	LAB(4096)	LAB(4096)	LAB(4096)	LAB(4096)	LAB(4096)
#Label	14	20	260	268	291	38
#Instance	2417	9963	4999	20770	19627	25000

Table 2. Detailed information of comparison methods. ✓ represent the method is able to handle the corresponding problem.

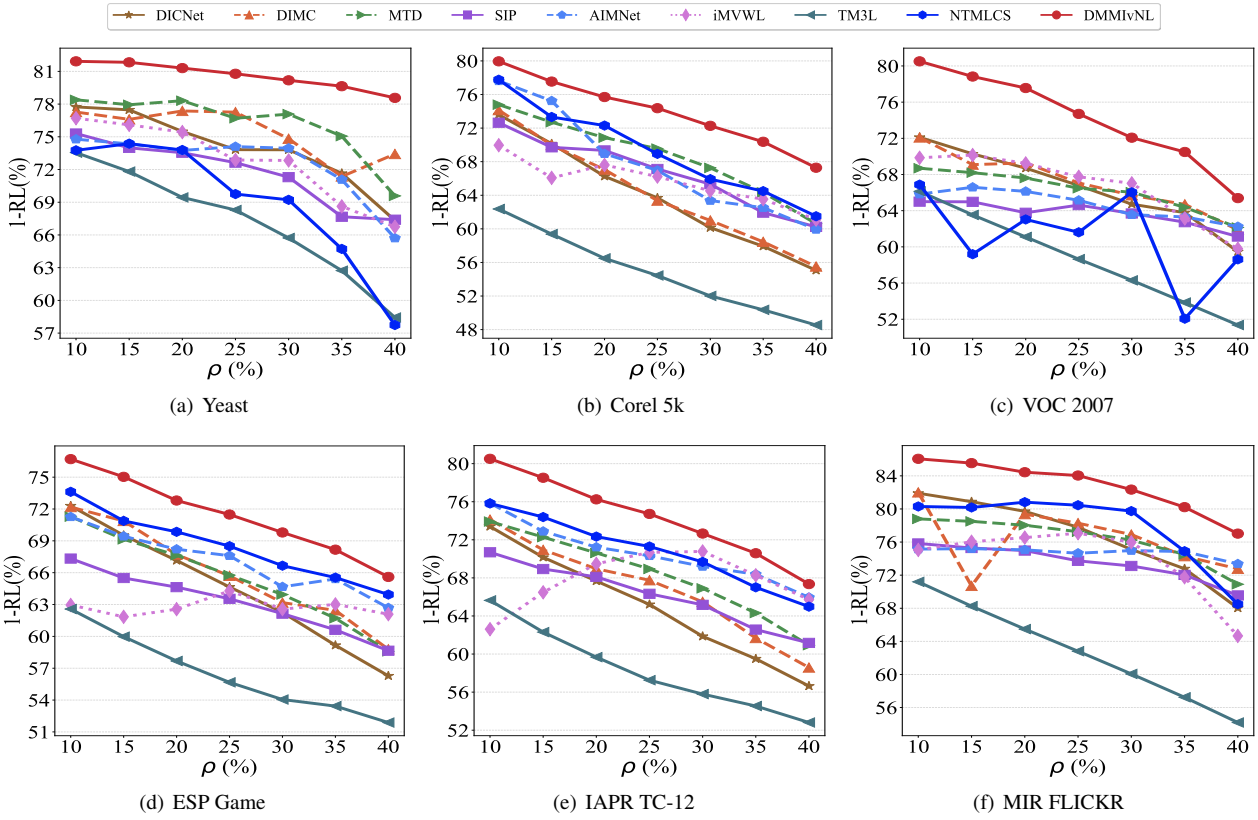
Method	Source	Year	Multi-label	Multi-view	Missing-view	Missing-label	Noisy-label
iMVWL	IJCAI	2018	✓	✓	✓	✓	
TM3L	ASC	2018	✓	✓	✓	✓	
ENTMLC	NeurIPS	2022	✓				✓
DICNet	AAAI	2023	✓	✓	✓	✓	
DIMC	TNNLS	2023	✓	✓	✓	✓	
AIMNet	AAAI	2024	✓	✓	✓	✓	
MTD	NeurIPS	2024	✓	✓	✓	✓	
SIP	ICML	2024	✓	✓	✓	✓	



(a) Yeast (b) Corel 5k (c) VOC 2007

(d) ESP Game (e) IAPR TC-12 (f) MIR FLICKR

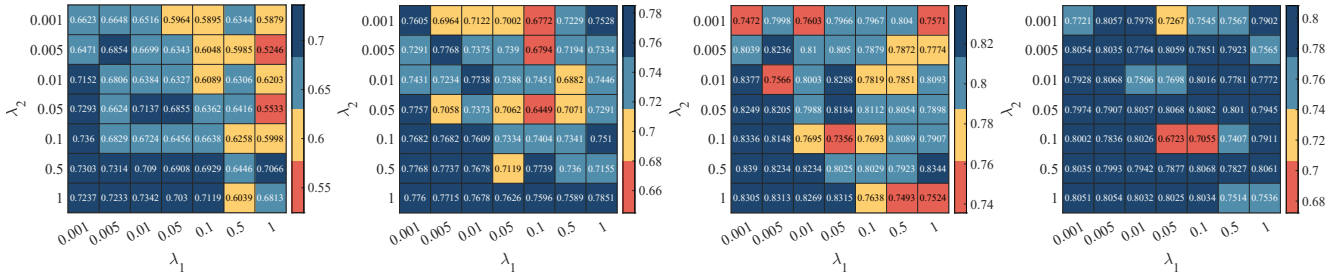
Figure 1. AP comparisons on six datasets with noise rate ρ varying from 10% to 40% while PER=50%.



(a) Yeast (b) Corel 5k (c) VOC 2007

(d) ESP Game (e) IAPR TC-12 (f) MIR FLICKR

Figure 2. RankingLoss comparisons on six datasets with noise rate ρ varying from 10% to 40% while PER=50%.



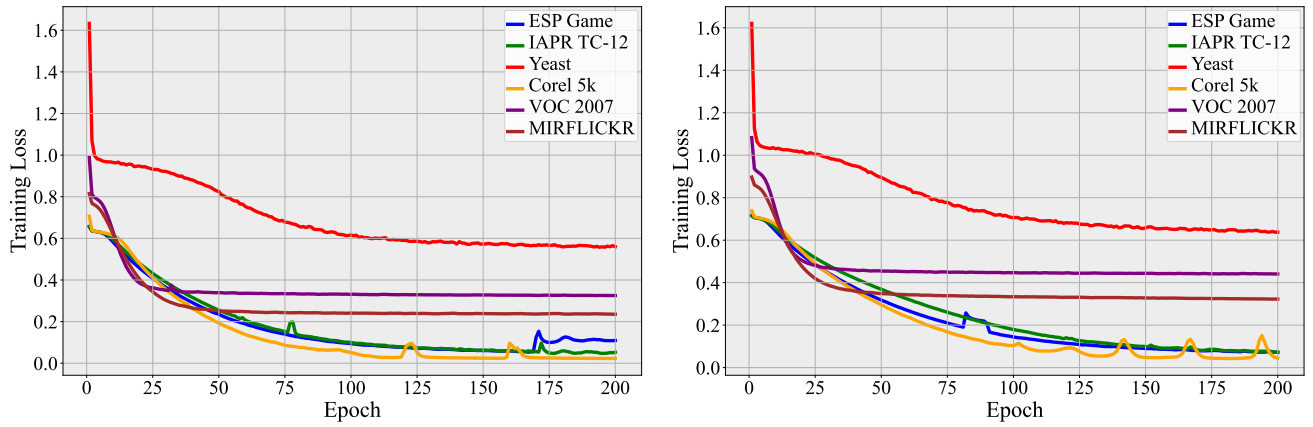
(a) Corel 5k (b) VOC 2007 (c) MIRFLICKR (d) Yeast

Figure 3. Parameter analysis of the trade-off parameters λ_1 and λ_2 on four datasets.

Table 3. Time efficiency during the training phase of the nine methods on three datasets. (Unit:second)

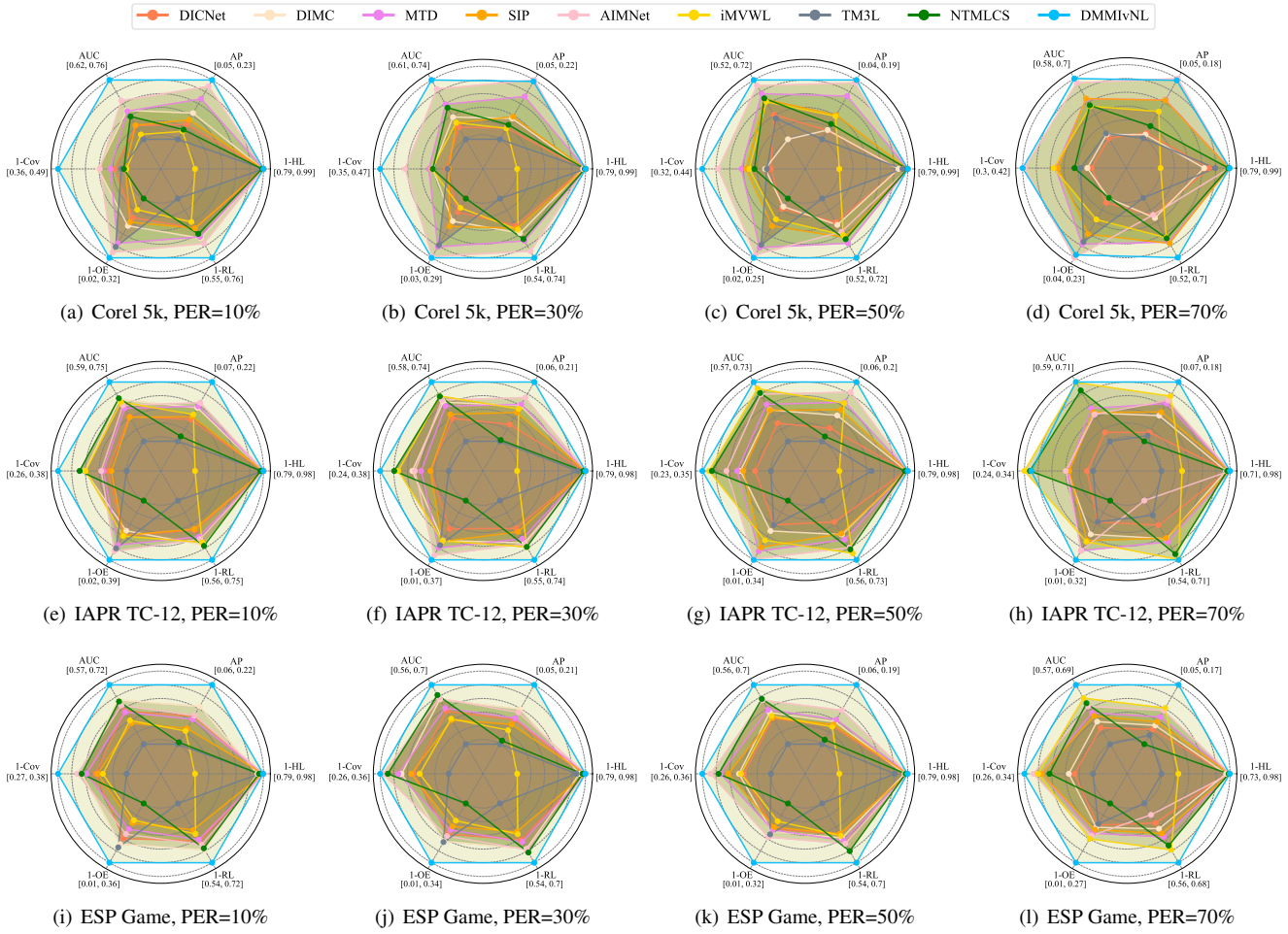
Data	DICNet	DIMC	MTD	SIP	AIMNet	iMVWL	TM3L	NTMLCS	DMMIvNL
Yeast	50.57	56.70	382.17	1170.41	1317.91	3.16	1.22	124.61	104.01
Corel 5k	723.57	786.74	2776.38	1718.96	1761.04	463.64	29.66	494.54	627.34
IAPR TC-12	3494.30	3937.17	5976.48	3915.21	3495.65	948.01	422.11	1915.92	2643.06

recognition, which contains 9963 images and 20 kinds of objects. **MIR FLICKR** consists of 25,000 images from the Flickr platform, annotated with a total of 38 tags. **Yeast** is a multi-label dataset containing 2417 images of yeast cells and each image is annotated with 14 labels indicating cellular characteristics. To validate the effectiveness of DMMIvNL, we compare it with eight state-of-the-art approaches, i.e., iMVWL [17], TM3L [21], DIMC [18], DICNet [9], MTD [11], AIMNet [10] SIP [12] and ENTMLC [6], which are discussed in detail in the Related Work section. We also provide a comprehensive overview of their sources and functions in Table 2.



(a) $\rho = 30\%$, PER=30% (b) $\rho = 40\%$, PER=50%

Figure 4. Convergence analysis on six datasets under different ρ and PER.

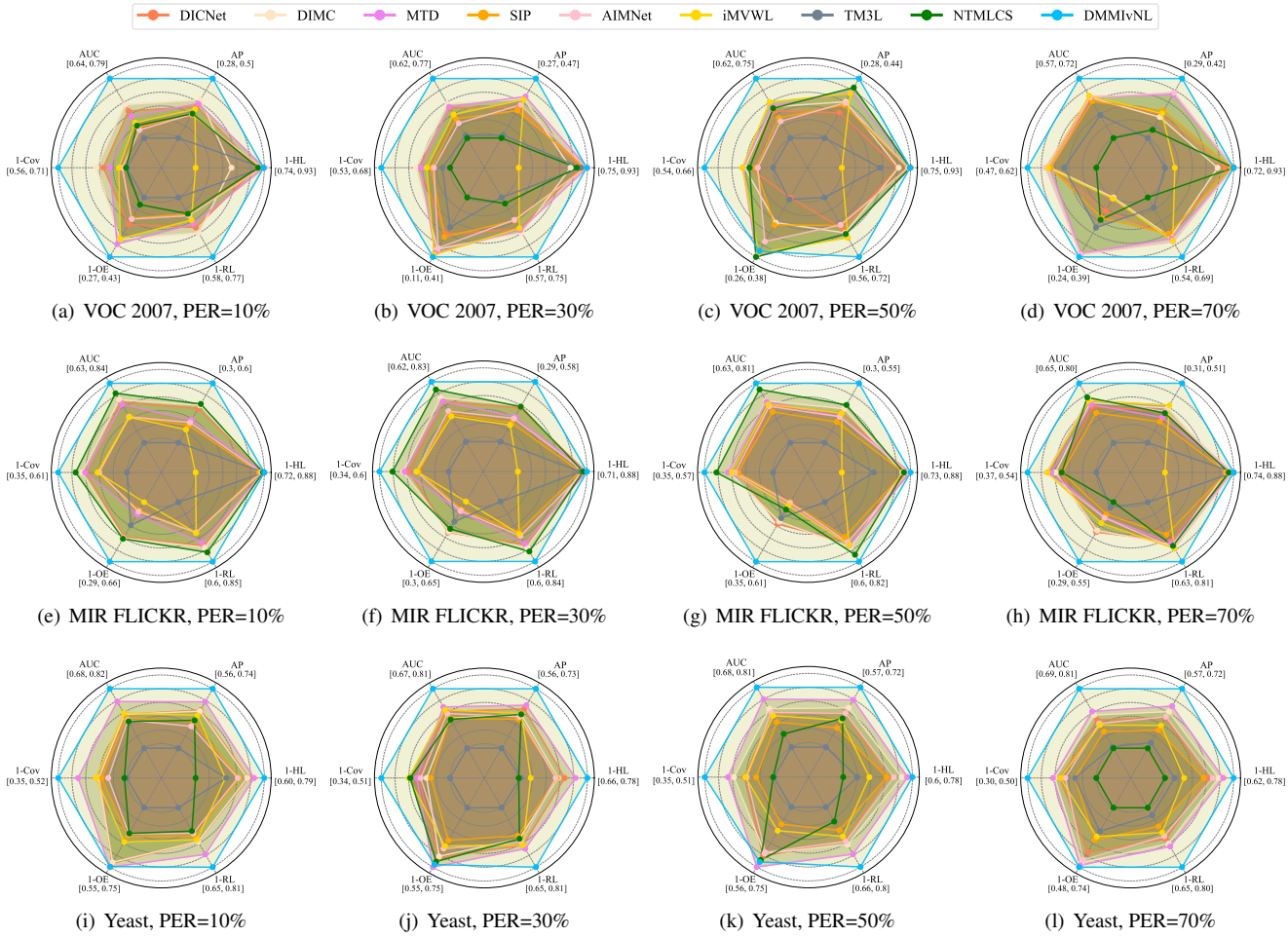


(a) Corel 5k, PER=10% (b) Corel 5k, PER=30% (c) Corel 5k, PER=50% (d) Corel 5k, PER=70%

(e) IAPR TC-12, PER=10% (f) IAPR TC-12, PER=30% (g) IAPR TC-12, PER=50% (h) IAPR TC-12, PER=70%

(i) ESP Game, PER=10% (j) ESP Game, PER=30% (k) ESP Game, PER=50% (l) ESP Game, PER=70%

Figure 5. Experimental results of nine methods on Corel 5k, IAPR TC-12 and ESP Game with PER varying from 10% to 70% while $\rho = 30\%$.



(a) VOC 2007, PER=10% (b) VOC 2007, PER=30% (c) VOC 2007, PER=50% (d) VOC 2007, PER=70%

(e) MIR FLICKR, PER=10% (f) MIR FLICKR, PER=30% (g) MIR FLICKR, PER=50% (h) MIR FLICKR, PER=70%

(i) Yeast, PER=10% (j) Yeast, PER=30% (k) Yeast, PER=50% (l) Yeast, PER=70%

Figure 6. Experimental results of nine methods on VOC 2007, MIR FLICKR and Yeast with PER varying from 10% to 70% while $\rho = 30\%$.

Implementation Details. Referring to multi-label learning works [7, 11], we employ Hamming Loss (HL), Ranking Loss (RL), OneError (OE), Coverage (Cov), Average Precision (AP), and Area Under Curve (AUC) as six metrics to unify the experimental standards. Higher AP and AUC values indicate better performance, while lower HL, RL, OE, and Cov values are preferred. Their evaluation contents are described below: 1) ACC measures the proportion of correctly predicted labels across all samples. 2) RL evaluates the accuracy of the model's ranking of predicted labels compared to true labels. 3) AP computes the area under the precision-recall curve, indicating the average precision achieved across all recall levels. 4) AUC quantifies the probability that a randomly selected positive instance is ranked higher by the model than a randomly selected negative instance across all possible threshold values. 5) OE evaluates whether the top-ranked label predicted by the model is incorrect. 6) Cov computes the number of labels the model needs to traverse to cover all true labels, reflecting the efficiency of the model's predicted label range. In our experiments, the two parameters λ_1 and λ_2 are selected in the range of $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$. Adam optimizer with the initial learning rate of 0.0001 is used for optimization of all datasets. All methods use the same dataset partition when conducting experiments, while the locations of view missing and label flipping are recorded and kept consistent.

B.2. Experiment Results

Comparative Experiment. To validate that our method can adapt to varying degrees of missing views and noisy labels, we conduct comparative experiments with ρ ranging from $\{10\%, 15\%, 20\%, 25\%, 30\%, 35\%, 40\%\}$ and PER selected from $\{10\%, 30\%, 50\%, 70\%\}$. Fig. 1 and 2 show AP and RankingLoss comparisons when noise rate increases, while Fig. 5 and 6 displays the metric distribution when

PER changes. It can be observed that regardless of the levels of view missing and noise rate, our method always occupies the top position in the line charts and the outermost edge of the radar charts, which indicates that DMMIvNL consistently outperforms the other eight methods. The results across all datasets and cases strongly validate that our method is both effective and robust for simultaneously handling insufficient features and incorrect annotations.

Parameter Determination, Convergence, and Time Efficiency Analysis. The parameters λ_1 and λ_2 are used to balance the roles of $\mathcal{L}_{ID}$ and $\mathcal{L}_R$ when handling noisy labels. From the heatmap shown in Fig. 3, the performance of DMMIvNL exhibits variability under different parameter setting. Besides, the optimal results are typically achieved when λ_1 falls within the range of (0.001, 0.005) and λ_2 within the range of (0.5, 1). Moreover, we present the loss variation trend under different ρ and PER in Fig. 4. The results show that our DMMIvNL demonstrates good convergence and gradually approaches the optimal network parameters under any conditions. Furthermore, the running times of all algorithms are reported in Table 3. From a time comparison perspective, our method is less time-consuming than most deep learning based methods. The reason may be that our main effort lies in developing loss functions with theoretical guarantees to adaptively address the dual issues of view missing and label noise. Thus, the constructed networks do not have a complex design and its structure primarily consists of fully connected layers. In conclusion, our method not only ensures stable performance but also comes with a resource expenditure that remains tolerable.

C. Visual Explanation

Visual of the Volume Minimization Network. To facilitate an intuitive understanding of the volume minimization network, we present visualizations in both the 2D plane and high-dimensional space. From the Fig. 7(a) and 8(a), we can see that in the sufficiently scattered assumption, condition 1 establishes a lower bound for the scatter of class-posterior probability points, i.e, there must exist a cone formed by a set of sample points that contains $\mathcal{R}$ (inner circle or sphere). Condition 2 imposes an upper limit on the dispersion of sample points, where they are enclosed by a cone formed by the columns of a permutation matrix $\boldsymbol{Q}$. The vertices of the cone are the anchor points with a probability of 1 belonging to a specific class. In the Fig. 7(b), the region encompassed by red lines is $\text{cone}\{\boldsymbol{H}^j\}$, which satisfies $\mathcal{R} \subseteq \text{cone}\{\boldsymbol{H}^j\} \subseteq \text{cone}\{\boldsymbol{Q}\}$. When the sufficiently scattered assumption is invalid shown in Fig. 7(c), the class-posterior probability points do not cover a sufficiently large area to provide adequate information for classification, making it difficult to distinguish between the categories. When the volume of the transition matrix for each category is independently optimized, as shown in Fig. 7(d) and 7(e), we can simultaneously obtain the optimal transition matrices and classifier, with the same concept conveyed in Fig. 8(b). However, due to the inherent label correlations in multi-label data, there is mutual influence among classes, which shows that the optimization process of each label's transition matrix is constrained by other labels.

Dominated by inter-label relationships, similar with [6], noisy multi-label data encompasses the noisy inner-class and inter-class transition matrices :

$$\boldsymbol{M}^{ij} = \begin{pmatrix} P(\bar{\boldsymbol{Y}}^i = 0 \mid \boldsymbol{Y}^j = 0) & P(\bar{\boldsymbol{Y}}^i = 0 \mid \boldsymbol{Y}^j = 1) \\ P(\bar{\boldsymbol{Y}}^i = 1 \mid \boldsymbol{Y}^j = 0) & P(\bar{\boldsymbol{Y}}^i = 1 \mid \boldsymbol{Y}^j = 1) \end{pmatrix},$$

where $\boldsymbol{M}^{ij}$ represents the noise transition matrix from label j to label i, with its elements primarily determined by the correlation between the two labels. If two labels like "fish" and "water" are strongly correlated, then since $P(\bar{\boldsymbol{Y}}^i = 1 \mid \boldsymbol{Y}^j = 1) = P(\bar{\boldsymbol{Y}}^i = 1; \boldsymbol{Y}^j = 1)/P(\boldsymbol{Y}^j = 1)$, a larger $P(\bar{\boldsymbol{Y}}^i = 1; \boldsymbol{Y}^j = 1)$ will have a considerable impact on $P(\bar{\boldsymbol{Y}}^i = 1 \mid \boldsymbol{Y}^j = 1)$. Moreover, through computations involving co-occurrence probabilities and conditional probabilities, we can derive a set of equations that establish the constraints between the inter-class transition matrices and the intra-class transition matrices:

$$\begin{cases} P\left(\bar{Y}^j = 0, \bar{Y}^i = 0\right) = P(Y^j = 0) T^j_{00} P(\bar{Y}^i = 0 | Y^j = 0) \\ \qquad\qquad + P(Y^j = 1) T^j_{01} P(\bar{Y}^i = 0 | Y^j = 1) \\ P\left(\bar{Y}^j = 0, \bar{Y}^i = 1\right) = P(Y^j = 0) T^j_{00} P(\bar{Y}^i = 1 | Y^j = 0) \\ \qquad\qquad + P(Y^j = 1) T^j_{01} P(\bar{Y}^i = 1 | Y^j = 1) \\ P\left(\bar{Y}^j = 1, \bar{Y}^i = 0\right) = P(Y^j = 0) T^j_{10} P(\bar{Y}^i = 0 | Y^j = 0) \\ \qquad\qquad + P(Y^j = 1) T^j_{11} P(\bar{Y}^i = 0 | Y^j = 1) \\ P\left(\bar{Y}^j = 1, \bar{Y}^i = 1\right) = P(Y^j = 0) T^j_{10} P(\bar{Y}^i = 1 | Y^j = 0) \\ \qquad\qquad + P(Y^j = 1) T^j_{11} P(\bar{Y}^i = 1 | Y^j = 1) \end{cases},$$

where the elements of the intra-class transition matrix $\boldsymbol{T}^j$ are jointly governed by the inter-class transition matrices $\boldsymbol{M}^{ij}$ and $\boldsymbol{M}^{ji}$. Therefore, the transition matrix for each category itself cannot be optimized completely independently, as it is influenced by the associated labels. As shown in Fig. 7(f), 7(g) and 8(c), label interactions compress the class-posterior probabilities of certain labels, thereby violating the sufficiently scattered assumption. As a result, minimizing the transition matrix volume alone is unlikely to function effectively in practice, which promotes the design of our cycle-consistency estimation framework. We leverage the volume maximization of the clean-class posterior to assist in the volume minimization of the transition matrix, which effectively avoids the estimation bias introduced by single-objective decision-making. Besides, by integrating forward and backward feedback mechanisms, it enhances noise identification robustness without relying on any prior assumptions.

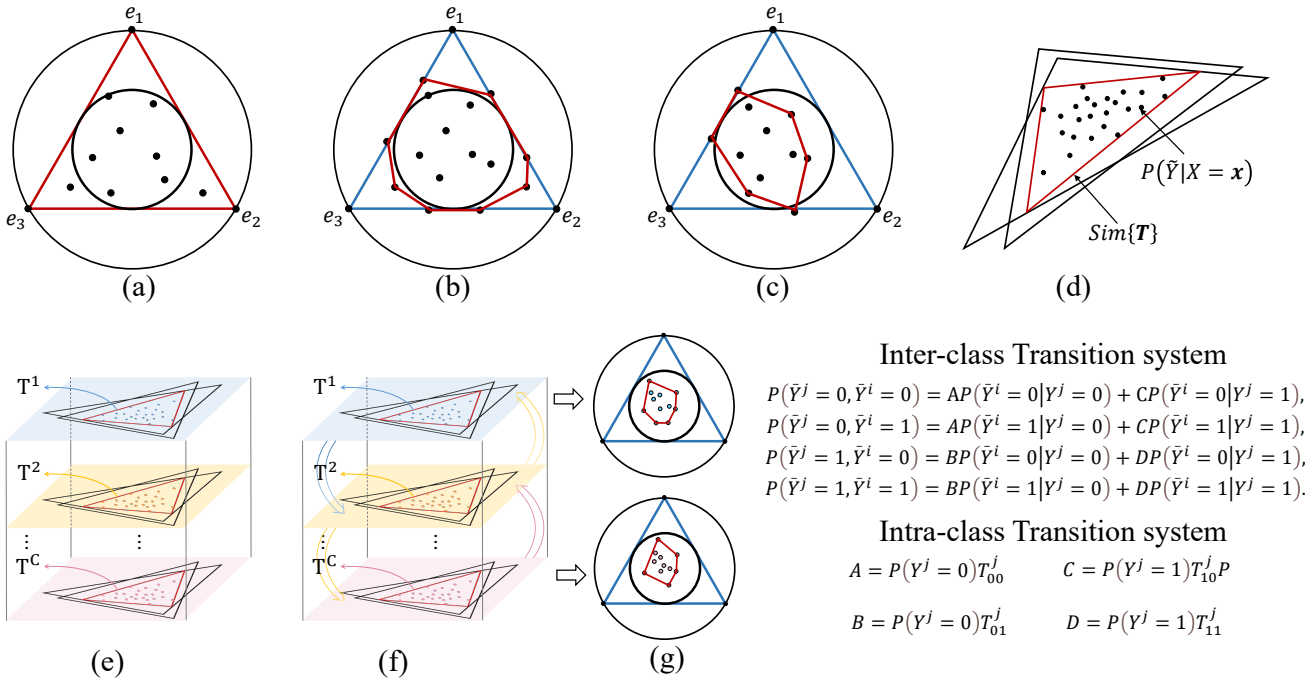


Figure 7. Illustration of the volume minimization network in the 2D plane. The black dots are class-posterior probability points; the inner circle is $\mathcal{R}\{\boldsymbol{v} \in \mathbb{R}^2 \mid \boldsymbol{v}^\top \mathbf{1} \geq \|\boldsymbol{v}\|_2\} \subseteq \text{cone}\{\boldsymbol{H}^j\}$; the vertices of the triangle are composed of anchor points; the region encompassed by red lines is $\text{cone}\{\boldsymbol{H}^j\}$. (a) identifies the transition matrix through anchor points; (b) shows the sufficiently scattered assumption; (c) highlights the failure of the sufficiently scattered assumption; (d) demonstrates the progressive optimization leading to the minimal volume transition matrix; (e) illustrates the state of independent optimization for the transition matrix of each label; (f) displays the mutual influence among classes driven by label correlations; (g) illustrates the failure of the sufficiently scattered assumption during the optimization process for $\boldsymbol{T}^1$ and $\boldsymbol{T}^C$ due to the influence of $\boldsymbol{T}^2$.

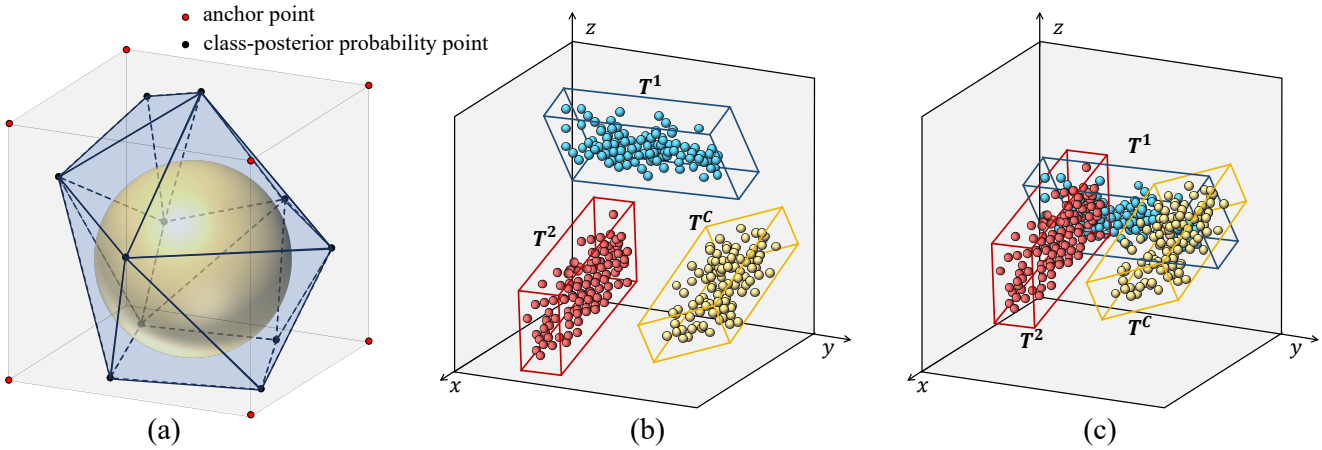


Figure 8. Illustration of the volume minimization network in the high-dimensional space. The cube is formed by anchor points; the inner sphere is $\mathcal{R}\{\boldsymbol{v} \in \mathbb{R}^2 \mid \boldsymbol{v}^\top \mathbf{1} \geq \|\boldsymbol{v}\|_2\} \subseteq \text{cone}\{\boldsymbol{H}^j\}$; the region fromed by black points is $\text{cone}\{\boldsymbol{H}^j\}$. (a) shows the sufficiently scattered assumption; (b) illustrates the state of independent optimization for the transition matrix of each label; (c) displays the mutual influence among classes driven by label correlations and the failure of the sufficiently scattered assumption during the optimization process after label interactions.

D. Theoretical Derivation and Proof

D.1. Complete Derivation of Information Bottleneck Theory Based Model

In this section, we give a detailed derivation about the upper bound of the problem (1):

$$\max \frac{1}{V}\sum_{v=1}^{V}\sum_{v^*\neq v}^{V}(I(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{z}^{v}|\boldsymbol{x}^{v}) - I(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v})) - \frac{\beta}{V}\sum_{v=1}^{V} I(\boldsymbol{z}^{v}|\boldsymbol{x}^{v};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}). \tag{1}$$

For the problem (1), we have the following equation by utilizing the definition of mutual information:

$$I(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{z}^{v}|\boldsymbol{x}^{v}) = \int\int p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v},\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\log\frac{p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v},\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})}{p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v})}d\boldsymbol{z}^{v^*}d\boldsymbol{z}^{v}. \tag{2}$$

Considering $p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*},\boldsymbol{z}^{v}|\boldsymbol{x}^{v}) = p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})$, we have

$$\begin{aligned} I(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{z}^{v}|\boldsymbol{x}^{v}) &= \int\int p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\log\frac{p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})}{p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^{v}/\boldsymbol{x}^{v})}d\boldsymbol{z}^{v^*}d\boldsymbol{z}^{v} \\ &= \int\int p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*}/\boldsymbol{z}^{v}|\boldsymbol{x}^{v})p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\log p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})d\boldsymbol{z}^{v^*}d\boldsymbol{z}^{v} \\ &+ \int\int p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v})\log\frac{1}{p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v})}d\boldsymbol{z}^{v^*}d\boldsymbol{z}^{v}. \end{aligned} \tag{3}$$

Since $H(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}) = -\int p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v})\log p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v})d\boldsymbol{z}^{v} \geq 0$, we have

$$\begin{aligned} I(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{z}^{v}|\boldsymbol{x}^{v}) &= \int\int p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\log p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})d\boldsymbol{z}^{v^*}d\boldsymbol{z}^{v} \\ &+ \int p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*}/\boldsymbol{z}^{v}|\boldsymbol{x}^{v})H(\boldsymbol{z}^{v}|\boldsymbol{x}^{v})d\boldsymbol{z}^{v^*} \\ &\geq \int\int p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\log p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})d\boldsymbol{z}^{v^*}d\boldsymbol{z}^{v} \\ &= \int\int p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\log q(\boldsymbol{z}^{v}|\boldsymbol{z}^{v^*})d\boldsymbol{z}^{v^*}d\boldsymbol{z}^{v} \\ &+ \int\int p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\log\frac{p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})}{q(\boldsymbol{z}^{v}|\boldsymbol{z}^{v^*})}d\boldsymbol{z}^{v^*}d\boldsymbol{z}^{v}. \end{aligned} \tag{4}$$

Based on the definition of the Kullback-Leibler divergence, we can get

$$D_{KL}(p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\|q(\boldsymbol{z}^{v}|\boldsymbol{z}^{v^*})) = \int p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\log\frac{p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})}{q(\boldsymbol{z}^{v}|\boldsymbol{z}^{v^*})}d\boldsymbol{z}^{v}. \tag{5}$$

Since $D_{KL}(p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\|q(\boldsymbol{z}^{v}|\boldsymbol{z}^{v^*})) \geq 0$, we have

$$\begin{aligned} I(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{z}^{v}|\boldsymbol{x}^{v}) &\geq \int\int p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\log q(\boldsymbol{z}^{v}|\boldsymbol{z}^{v^*})d\boldsymbol{z}^{v^*}d\boldsymbol{z}^{v} \\ &+ \int p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})D_{KL}(p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\|q(\boldsymbol{z}^{v}|\boldsymbol{z}^{v^*}))d\boldsymbol{z}^{v^*} \\ &\geq \int\int p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v}/\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\log q(\boldsymbol{z}^{v}|\boldsymbol{z}^{v^*})d\boldsymbol{z}^{v^*}d\boldsymbol{z}^{v} \\ &= \int\int p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v},\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})\log q(\boldsymbol{z}^{v}|\boldsymbol{z}^{v^*})d\boldsymbol{z}^{v^*}d\boldsymbol{z}^{v}. \end{aligned} \tag{6}$$

For the second term $I(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v})$, we start from the definition and obtain the following upper bound:

$$
\begin{aligned}
I(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) &= \int\int p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \log \frac{p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v})}{p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v})} d\boldsymbol{s}^{v^*} d\boldsymbol{s}^{v} \\
&= \int\int p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \log \frac{p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v}/\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*})}{p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v})} d\boldsymbol{s}^{v^*} d\boldsymbol{s}^{v} \\
&= \int\int p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \log \frac{p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v}/\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*})}{q(\boldsymbol{s}^{v}|\boldsymbol{s}^{v^*})} d\boldsymbol{s}^{v^*} d\boldsymbol{s}^{v} \\
&+ \int\int p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \log \frac{q(\boldsymbol{s}^{v}|\boldsymbol{s}^{v^*})}{p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v})} d\boldsymbol{s}^{v^*} d\boldsymbol{s}^{v} \\
&= \int\int p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \log \frac{p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v}/\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*})}{q(\boldsymbol{s}^{v}|\boldsymbol{s}^{v^*})} d\boldsymbol{s}^{v^*} d\boldsymbol{s}^{v} \\
&- \int p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*}/\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \int p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \log \frac{p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v})}{q(\boldsymbol{s}^{v}|\boldsymbol{s}^{v^*})} d\boldsymbol{s}^{v^*} d\boldsymbol{s}^{v} \\
&= \int\int p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \log \frac{p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v}/\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*})}{q(\boldsymbol{s}^{v}|\boldsymbol{s}^{v^*})} d\boldsymbol{s}^{v^*} d\boldsymbol{s}^{v} \\
&- \int p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*}/\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) D_{KL}(p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \| q(\boldsymbol{s}^{v}|\boldsymbol{s}^{v^*})) d\boldsymbol{s}^{v^*} \\
&\leq \int\int p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \log \frac{p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v}/\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*})}{q(\boldsymbol{s}^{v}|\boldsymbol{s}^{v^*})} d\boldsymbol{s}^{v^*} d\boldsymbol{s}^{v} \\
&= \int\int p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \log \frac{p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v})}{p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*})q(\boldsymbol{s}^{v}|\boldsymbol{s}^{v^*})} d\boldsymbol{s}^{v^*} d\boldsymbol{s}^{v} \\
&= D_{KL}(p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \| p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*})q(\boldsymbol{s}^{v}|\boldsymbol{s}^{v^*})).
\end{aligned}
\tag{7}
$$

Similarly, we can derive the upper bound for the last term, which exhibit structure analogous to that of Eq. (7):

$$
I(\boldsymbol{z}^{v}|\boldsymbol{x}^{v};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \leq D_{KL}(p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v};\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \| p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v})q(\boldsymbol{z}^{v}|\boldsymbol{s}^{v})). \tag{8}
$$

By combining Eqs. (6), (7), and (8), the objective of our method is naturally transformed into minimizing its upper bound:

$$
\begin{aligned}
\mathcal{L}_{IB} = \frac{1}{V}\sum_{v=1}^{V}\sum_{v^*\neq v}^{V} \Big[&-D_{KL}^{\dagger}(p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v},\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*}) \| q(\boldsymbol{z}^{v}|\boldsymbol{z}^{v^*})) \\
&+ D_{KL}(p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*},\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \| p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*})q(\boldsymbol{s}^{v}|\boldsymbol{s}^{v^*})) \Big] \\
&+ \frac{\beta}{V}\sum_{v=1}^{V} D_{KL}(p(\boldsymbol{z}^{v}|\boldsymbol{x}^{v},\boldsymbol{s}^{v}|\boldsymbol{x}^{v}) \| p(\boldsymbol{s}^{v}|\boldsymbol{x}^{v})q(\boldsymbol{z}^{v}|\boldsymbol{s}^{v})).
\end{aligned}
\tag{9}
$$

Each term in Eq. (9) denoted as $D_{KL}^{\dagger(\boldsymbol{z}^{v},\boldsymbol{z}^{v^*})}$, $D_{KL}^{(\boldsymbol{s}^{v},\boldsymbol{s}^{v^*})}$ and $D_{KL}^{(\boldsymbol{s}^{v},\boldsymbol{z}^{v})}$, then we can obtain

$$
\mathcal{L}_{IB} = \frac{1}{V}\sum_{v=1}^{V}\sum_{v^*\neq v}^{V}(-D_{KL}^{\dagger(\boldsymbol{z}^{v},\boldsymbol{z}^{v^*})} + D_{KL}^{(\boldsymbol{s}^{v},\boldsymbol{s}^{v^*})}) + \frac{\beta}{V}\sum_{v=1}^{V} D_{KL}^{(\boldsymbol{s}^{v},\boldsymbol{z}^{v})}. \tag{10}
$$

D.2. Proof of the Theorem 1

The sufficiently scattered assumption and following lemmas are essential prerequisite for our proof.

Definition 1. *(Sufficiently Scattered [2]). The clean class-posterior probability $P(\boldsymbol{Y}^{j}|\boldsymbol{X})$ is said to be sufficiently scattered only if there exists a feature set $\mathcal{H} = \{\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_m\}$ such that the matrix $\boldsymbol{H}^{j} = [P(\boldsymbol{Y}^{j}|\boldsymbol{X} = \hat{\boldsymbol{x}}_1), \ldots, P(\boldsymbol{Y}^{j}|\boldsymbol{X} = \hat{\boldsymbol{x}}_m)]$ satisfies the following conditions: 1) $\mathcal{R} \subseteq \mathrm{cone}\{\boldsymbol{H}^{j}\}$, where $\mathcal{R} = \{\boldsymbol{v} \in \mathbb{R}^2 \mid \boldsymbol{v}^{\top}\boldsymbol{1} \geq \|\boldsymbol{v}\|_2\}$ and $\mathrm{cone}\{\boldsymbol{H}^{j}\}$ denotes the convex cone formed by columns of $\boldsymbol{H}^{j}$. 2) $\mathrm{cone}\{\boldsymbol{H}^{j}\} \nsubseteq \mathrm{cone}\{\boldsymbol{Q}\}$, where $\boldsymbol{Q} \in \mathbb{R}^{2\times 2}$ is any unitary matrix that is not a permutation matrix.*

Lemma 1. *If $\mathcal{K}_1$ and $\mathcal{K}_2$ are convex cones, and $\mathcal{K}_1 \in \mathcal{K}_2$, then $dual\{K_2\} \in dual\{K_1\}$.*

Lemma 2. *If $\boldsymbol{A}$ is an invertible matrix, then $dual(\boldsymbol{A}) = \text{cone}(\boldsymbol{A}^{-\top})$.*

Our goal is to demonstrate that solving the following problem can simultaneously yield the optimal transition matrix for each category and classifier:

$$\begin{cases} \min\limits_{\hat{\boldsymbol{T}}^j} \sum\limits_{j=1}^{C} \text{vol}(\hat{\boldsymbol{T}}^j) \\ s.t.\ \hat{\boldsymbol{T}}^j \begin{bmatrix} h_\theta(\hat{\boldsymbol{x}})_j \\ 1 - h_\theta(\hat{\boldsymbol{x}})_j \end{bmatrix} = P(\bar{\boldsymbol{Y}}^j | X = \hat{\boldsymbol{x}}), \end{cases} \tag{11}$$

where $\text{vol}(\hat{\boldsymbol{T}}^j)$ denotes a measure to the volume of $\text{Sim}\{\hat{\boldsymbol{T}}^j\}$ and we choose a common choice that $\text{vol}(\hat{\boldsymbol{T}}^j) = \det(\hat{\boldsymbol{T}}^j)$. Then under the condition that the clean class-posterior satisfies the sufficiently scattered assumption, we can theoretically establish addressing the problem (11) will drive $\hat{\boldsymbol{T}}^j$ to converge to $\boldsymbol{T}^j$, while $h_\theta(\hat{\boldsymbol{x}})$ accurately approximates $P(\boldsymbol{Y}|\boldsymbol{X} = \hat{\boldsymbol{x}})$:

Theorem 1. *If each $P(\boldsymbol{Y}^j|\boldsymbol{X})$ is sufficiently scattered, then $\boldsymbol{T}_\star^j = \boldsymbol{T}^j$ and $(h_{\theta_\star}(\hat{\boldsymbol{x}})_j, 1 - h_{\theta_\star}(\hat{\boldsymbol{x}})_j)^T = P(\boldsymbol{Y}^j|\boldsymbol{X} = \hat{\boldsymbol{x}})$ $(j = 1, 2 \cdots C)$ must hold, where $(\boldsymbol{T}_\star^1, \boldsymbol{T}_\star^2, \cdots, \boldsymbol{T}_\star^C, \theta_\star)$ are the optimal solutions of the problem (11).*

Proof. $(\hat{\boldsymbol{T}}_\star^1, \hat{\boldsymbol{T}}_\star^2, \cdots, \hat{\boldsymbol{T}}_\star^C, \theta_\star)$ is denoted as a feasible solution of the problem (11), i.e.,

$$\hat{\boldsymbol{T}}_\star^j (h_{\theta_\star}(\hat{\boldsymbol{x}})_j, 1 - h_{\theta_\star}(\hat{\boldsymbol{x}})_j)^T = \boldsymbol{T}^j P(\boldsymbol{Y}^j|\boldsymbol{X}) = P(\bar{\boldsymbol{Y}}^j|\boldsymbol{X}). \tag{12}$$

According to the sufficiently scattered assumption, we have the matrix $\boldsymbol{H}^j = [P(\boldsymbol{Y}^j|\boldsymbol{X} = \hat{\boldsymbol{x}}_1), \ldots, P(\boldsymbol{Y}^j|\boldsymbol{X} = \hat{\boldsymbol{x}}_m)]$ defined on the set $\mathcal{H} = \{\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_m\}$. Based on the output of the classifier, the matrix $\boldsymbol{H}^j$ is expressed as below:

$$\boldsymbol{H}_\star^j = \begin{bmatrix} h_{\theta_\star}(\hat{\boldsymbol{x}}_1)_j & \ldots & h_{\theta_\star}(\hat{\boldsymbol{x}}_m)_j \\ 1 - h_{\theta_\star}(\hat{\boldsymbol{x}}_1)_j & \ldots & 1 - h_{\theta_\star}(\hat{\boldsymbol{x}}_m)_j \end{bmatrix}. \tag{13}$$

It follows that $\boldsymbol{T}_\star^j \boldsymbol{H}_\star^j = \boldsymbol{T}^j \boldsymbol{H}^j$ holds. Since $P(\bar{\boldsymbol{Y}}^j = 0 \mid \boldsymbol{Y}^j = 1) + P(\bar{\boldsymbol{Y}}^j = 1 \mid \boldsymbol{Y}^j = 0) < 1$, both $\boldsymbol{T}_\star^j$ and $\boldsymbol{T}^j$ have full rank. Therefore, there exists an invertible matrix $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ such that $\boldsymbol{T}_\star^j = \boldsymbol{T}^j \boldsymbol{A}^{-\top}$, where $\boldsymbol{A}^{-\top} = \boldsymbol{H}^j \boldsymbol{H}_\star^{j\dagger}$ and $\boldsymbol{H}_\star^{j\dagger} = \boldsymbol{H}_\star^{j\top} (\boldsymbol{H}_\star^j \boldsymbol{H}_\star^{j\top})^{-1}$. Since $\mathbf{1}^\top \boldsymbol{H}^j = \mathbf{1}^\top$ and $\mathbf{1}^\top \boldsymbol{H}_\star^j = \mathbf{1}^\top$, we can obtain

$$\mathbf{1}^\top \boldsymbol{A}^{-\top} = \mathbf{1}^\top \boldsymbol{H}^j \boldsymbol{H}_\star^{j\dagger} = \mathbf{1}^\top \boldsymbol{H}_\star^{j\dagger} = \mathbf{1}^\top \boldsymbol{H}_\star^j \boldsymbol{H}_\star^{j\dagger} = \mathbf{1}^\top. \tag{14}$$

Let $\boldsymbol{v} \in \text{cone}\{\boldsymbol{H}^j\}$, i.e., $\boldsymbol{v} = \boldsymbol{H}^j \boldsymbol{u}$, where $\boldsymbol{u} \geq 0$ and $\boldsymbol{u} \in \mathbb{R}^{m\times 1}$. Since $\boldsymbol{H}^j = \boldsymbol{A}^{-\top} \boldsymbol{H}_\star^j$, $\boldsymbol{v}$ can be expressed as $\boldsymbol{v} = \boldsymbol{A}^{-\top} \tilde{\boldsymbol{u}}$ where $\tilde{\boldsymbol{u}} = \boldsymbol{H}_\star^j \boldsymbol{u} \geq 0$, which implies that $\boldsymbol{v} \in \text{cone}\{\boldsymbol{A}^{-\top}\}$. Thus, based on the recursive relationship, we can deduce that $\text{cone}\{\boldsymbol{H}^j\} \in \text{cone}\{\boldsymbol{A}^{-\top}\}$.

By applying condition 1 of the sufficiently scattered assumption, we obtain that $\mathcal{R} \subseteq \text{cone}\{\boldsymbol{H}^j\} \subseteq \text{cone}\{\boldsymbol{A}^{-\top}\}$, where $\mathcal{R} = \{\boldsymbol{v} \in \mathbb{R}^2 \mid \boldsymbol{v}^\top \mathbf{1} \geq \sqrt{C} - 1 \|\boldsymbol{v}\|_2\}$. Using Lemmas 1 and 2, we have $\text{cone}\{\boldsymbol{A}\} \subseteq dual\{\mathcal{R}\}$, where $dual\{\mathcal{R}\} = \{\boldsymbol{v} \in \mathbb{R}^2 | \|\boldsymbol{v}\|_2 \leq \mathbf{1}^\top \boldsymbol{v}\}$ is the dual cone of $\mathcal{R}$. Then we can derive the following inequalities:

$$|det(\boldsymbol{A})| \overset{①}{\leq} \prod_{i=1}^{2} \|\boldsymbol{A}_{:,i}\|_2 \overset{②}{\leq} \prod_{i=1}^{2} \mathbf{1}^\top \boldsymbol{A}_{:,i} \overset{③}{\leq} (\frac{\sum_{i=1}^{2} \mathbf{1}^\top \boldsymbol{A}_{:,i}}{2})^2 = (\frac{\mathbf{1}^\top \boldsymbol{A} \mathbf{1}}{2})^2 \overset{④}{=} 1, \tag{15}$$

where ① is by the Hadamard's inequality, ② is by $\text{cone}\{\boldsymbol{A}\} \subseteq dual\{\mathcal{R}\}$, ③ is by the arithmetic-geometric mean inequality, ④ is by Eq. (14). Since $|det(\boldsymbol{A})|^{-1} = |det(\boldsymbol{A}^{-\top})|$ and $det(\boldsymbol{T}_\star^j) = det(\boldsymbol{T}^j)|det(\boldsymbol{A})|^{-1}$, we can get $det(\boldsymbol{T}_\star^j) \geq det(\boldsymbol{T}^j)$. In the original optimization problem (11), when $\min_{\hat{\boldsymbol{T}}^j} \sum_j \text{vol}(\hat{\boldsymbol{T}}^j)$ achieves its minimum, each $\text{vol}(\boldsymbol{T}^j)$ simultaneously reaches its individual minimum. Leveraging the relationship $det(\boldsymbol{T}^j) = \text{vol}(\boldsymbol{T}^j)$, it follows that $det(\boldsymbol{T}_\star^j) \leq det(\boldsymbol{T}^j)$. Therefore, we can get the conclusion:

$$det(\boldsymbol{T}_\star^j) = det(\boldsymbol{T}^j). \tag{16}$$

The equality in ① holds only if $\boldsymbol{A}$ is a column-orthogonal matrix with each column having the same magnitude. Consequently, $\boldsymbol{A}^{-\top}$ also satisfies this condition. The elements of $\boldsymbol{A}^{-\top}$ are denoted as a_{00}, a_{01}, a_{10}, and a_{11}. Combining with Eq.

(14), the following system of equations can be derived:

$$\begin{cases} a_{00}^2 + a_{10}^2 = a_{01}^2 + a_{11}^2 \\ a_{00}a_{01} + a_{10}a_{11} = 0 \\ a_{00} + a_{10} = a_{01} + a_{11} = 1 \end{cases}. \tag{17}$$

Therefore, we can deduce that $\boldsymbol{A}$ is an identity matrix and $\boldsymbol{T}_\star^j = \boldsymbol{T}^j$, which also makes $\text{cone}\{\boldsymbol{H}^j\} \in \text{cone}\{\boldsymbol{A}^{-\top}\}$ satisfy the condition 2 of the sufficiently scattered assumption. Then from Eq. (12), we derive that $(h_{\theta_\star}(\hat{\boldsymbol{x}})_j, 1 - h_{\theta_\star}(\hat{\boldsymbol{x}})_j)^T = P(\boldsymbol{Y}^j|\boldsymbol{X})$. Finally, $(\boldsymbol{T}_\star^j, (h_{\theta_\star}(\hat{\boldsymbol{x}})_j, 1 - h_{\theta_\star}(\hat{\boldsymbol{x}})_j)^T)$ $(j = 1, 2, \cdots, C)$ are proved as the unique optimal solutions of the problem (11). □

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China [Grant No. 62376281], and the Key NSF of China under Grant No. 62136005.

References

[1] Jia-Yao Chen, Shao-Yuan Li, Sheng-Jun Huang, and Chm. Unm: A universal approach for noisen, songcan and wang, lei and xie, ming-kun. *IEEE Transactions on Knowledge and Data Engineering*, 36:4968–4980, 2024. 1

[2] Xiao Fu, Kejun Huang, and Nicholas D Sidiropoulos. On identifiability of nonnegative matrix factorization. *IEEE Signal Processing Letters*, 25(3):328–332, 2018. 9

[3] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017. 1

[4] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *ICCV*, pages 101–110, 2019. 1

[5] Youngdong Kim, Juseung Yun, Hyounguk Shon, and Junmo Kim. Joint negative and positive learning for noisy labels. In *CVPR*, pages 9442–9451, 2021. 1

[6] Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. Estimating noise transition matrix with label correlations for noisy multi-label learning. In *NeurIPS*, pages 24184–24198, 2022. 1, 3, 6

[7] Xiang Li and Songcan Chen. A concise yet effective model for non-aligned incomplete multi-view and missing multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5918–5932, 2021. 1, 5

[8] Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. In *ICML*, pages 6403–6413, 2021. 1

[9] Chengliang Liu, Jie Wen, Xiaoling Luo, Chao Huang, Zhihao Wu, and Yong Xu. Dicnet: Deep instance-level contrastive network for double incomplete multi-view multi-label classification. In *AAAI*, pages 8807–8815, 2023. 1, 3

[10] Chengliang Liu, Jinlong Jia, Jie Wen, Yabo Liu, Xiaoling Luo, Chao Huang, and Yong Xu. Attention-induced embedding imputation for incomplete multi-view partial multi-label classification. In *AAAI*, pages 13864–13872, 2024. 1, 3

[11] Chengliang Liu, Jie Wen, Yabo Liu, Chao Huang, Zhihao Wu, Xiaoling Luo, and Yong Xu. Masked two-channel decoupling framework for incomplete multi-view weak multi-label learning. In *NeurIPS*, 2024. 1, 3, 5

[12] Chengliang Liu, Gehui Xu, Jie Wen, Yabo Liu, Chao Huang, and Yong Xu. Partial multi-view multi-label classification via semantic invariance learning and prototype modeling. In *ICML*, 2024. 1, 3

[13] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *ICML*, pages 6543–6553. PMLR, 2020. 1

[14] Shilong Ou, Zhe Xue, Yawen Li, Meiyu Liang, Yuanqiang Cai, and Junjiang Wu. View-category interactive sharing transformer for incomplete multi-view multi-label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27467–27476, 2024. 1

[15] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 1944–1952, 2017. 1

[16] Feng Sun, Ming-Kun Xie, and Sheng-Jun Huang. A deep model for partial multi-label image classification with curriculum-based disambiguation. *Machine Intelligence Research*, pages 1–14, 2024. 1

[17] Qiaoyu Tan, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang. Incomplete multi-view weak-label learning. In *IJCAI*, pages 2703–2709, 2018. 1, 3

[18] Jie Wen, Chengliang Liu, Shijie Deng, Yicheng Liu, Lunke Fei, Ke Yan, and Yong Xu. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8): 11396–11408, 2023. 1, 3

[19] Ming-Kun Xie and Sheng-Jun Huang. Ccmn: A general framework for learning with class-conditional multi-label noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):154–166, 2022. 1

[20] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *NIPS*, 31, 2018. 1

[21] Dawei Zhao, Qingwei Gao, Yixiang Lu, and Dong Sun. Two-step multi-view and multi-label learning with missing label via subspace learning. *Applied Soft Computing*, 102:107120, 2021. 1, 3